OPTIMAL CLUSTERING IN MULTI-STAGE SAMPLES

Robert Clark, Australian Bureau of Statistics
PO Box 10, ACT 2616 AUSTRALIA; email robert.clark@abs.gov.au

Key Words: Sample survey, Multi-Stage sample design

Abstract

In two-stage sampling, the sample numbers of primary and secondary sampling units are often chosen to minimize cost and variance. This problem can be referred to as optimal clustering, as the relative first and second stage sample sizes control the degree of geographical clustering, for area-based surveys. In this article, I discuss the existing theory for optimal clustering of area-based surveys and its application to a recent Redesign of the Australian Monthly Labour Force Survey. Telephone interviewing was recently introduced to the survey, resulting in a smaller and less clustered optimal design than for face to face interviewing.

1. Introduction

This paper discusses the allocation of sampling resources to two stages of an area-based sample. Typically, the first stage will be a sample of some regions or areas (primary sampling units or PSUs), while the second stage will be a sample of persons or dwellings within these areas. The following notation will be used to describe the problem:

m = number of first stage or PSUs in the sample;
q = number of second stage or final units selected in each PSU (sometimes referred to as the optimal cluster size);
n = mq = total number of final units in sample.

It will be assumed that PSUs are selected by stratified probability proportional to size or stratified random sampling. In principle, the parameter m could be chosen separately for each stratum in a stratified design, while the parameter q could be chosen separately for each stratum or even for each PSU.

There is some literature on the optimal choice of q and m to reduce both cost and variance. Models can be formulated for the total cost and the variance of one or more variables, in terms of q and m. The parameters q and m can then be chosen to minimize cost for one or more variance constraints; the variance can also be minimized for fixed cost. Section 2 summarizes the existing literature which includes: several forms of cost and variance models and the resulting q and m; the estimation of optimal cluster size from a pilot sample; optimal clustering for multiple variance constraints; and calculating different optimal cluster sizes for different area types.

In Section 3, the application of this theory to the Australian Monthly Labour Force Survey (MLFS) is described. This survey has previously been conducted by personal interviewing (PI), but telephone interviewing (TI) has recently been introduced. It was found that the optimal TI design was much less clustered than the optimal PI design, containing 9.2% fewer dwellings and 44% more PSUs than the optimal design under personal interviewing. The optimal clustering for the MLFS was complicated by a constraint of proportional sampling within state, and the need for adequate state as well as national precision. The calculation of the cost and variance models is also discussed.

2. Methods for Setting a Single Cluster Size

Hansen, Hurwitz and Madow (1953) (chap.9) contains a discussion of optimal clustering for quite general cost models. The following variance model was assumed:

$$\textbf{(1)} \qquad V = V_1 + V_2 m^{-1} + V_3 m^{-1} q^{-1};$$

while two possible cost models were discussed for two stage designs:

$$\textbf{(2)} \qquad C = C_1\sqrt{m} + C_2 m + C_3 mq \text{ ; and}$$
$$C = C_1\sqrt{m} + C_2 m + C_3 mq + C_4 m\sqrt{q} \ .$$

where C = total cost, V = variance of the estimator of interest, and C_i and V_i are parameters estimated in some way. There was no algebraic solution for the values of q and m minimizing cost for fixed variance, however algorithms were provided. A cost model for three stage designs was also given; this was a generalization of the second cost formula above.

Cochran(1977) (chap.10) and Snedecor and Cochran (1980) (chap.21) used the variance model (1), but a simpler cost model:

$$\textbf{(3)} \qquad C = C_1 + C_2 m + C_3 mq$$

Typically, C_1 would represent setup costs including interviewer overheads; C_2 would represent travel between PSUs; and C_3 would represent travel within PSUs and interviewing time. For this formulation of the cost and variance models, the values of m and q which minimize the cost C for fixed variance V=K are:

(4)
$$m_{opt} = (K - V_1)^{-1}(V_2 + V_2^{\frac{1}{2}}V_3^{\frac{1}{2}}C_2^{-\frac{1}{2}}C_3^{\frac{1}{2}})$$
$$q_{opt} = V_2^{-\frac{1}{2}}V_3^{\frac{1}{2}}C_2^{\frac{1}{2}}C_3^{-\frac{1}{2}}$$
$$n_{opt} = (K - V_1)^{-1}(V_2^{\frac{1}{2}}V_3^{\frac{1}{2}}C_2^{\frac{1}{2}}C_3^{-\frac{1}{2}} + V_3).$$

In fact, Cochran(1977) noted that $q=q_{opt}$ also minimizes: the product of cost and variance with no constraint; and variance for fixed cost. Clearly different values for m and n than those given in (4) would be optimal for these problems.

It is notable that in formula (4), q_{opt} depends on the ratios $\frac{V_3}{V_2}$ and $\frac{C_2}{C_3}$, not on any other cost or variance parameters, so that estimating these ratios accurately should be the main aim of cost and variance modelling. Equivalently, it is the intra-PSU correlation $\rho = \frac{V_2}{V_2+V_3}$ which is of interest from the variance model.

Brooks(1955) discussed the estimation from a pilot sample of the optimal cluster size q_{opt} as given by (4). Cost and variance models (1) and (3) were assumed, with $C_1=V_1=0$. It was further assumed that the cost model was exact, however V_2 and V_3 would be estimated from a pilot sample. It was noted that (4) would normally give a non-integer value for q_{opt}. This could be achieved by assigning the truncation and ceiling of q_{opt} to appropriate proportions of PSUs, however this would be inconvenient in practice and provide negligible gains. Instead, a rounding procedure was suggested, where an integer value q_{int} is chosen to satisfy:

$$q_{int}(q_{int} - 1) \le q_{opt}^2 \le q_{int}(q_{int} + 1).$$

To assist with designing a pilot test to calculate the optimal cluster size, Brooks examined different sizes of pilot test (first and second stage) and different values of V_2/V_3, based on a two-level normal model. The pilot designs providing given levels of precision for q_{int} were tabulated.

Waters and Chester (1987) discussed the minimization of cost subject to variance constraints for several variables. This problem can be written as:

(5) minimize $C = C_1 + C_2m + C_3mq$, with respect to m,q;

subject to $V^{(s)} = V_1^{(s)} + V_2^{(s)}m^{-1} + V_3^{(s)}m^{-1}q^{-1} \le K^{(s)}$ for s=1, ..., S;
where $V^{(s)}$ are the variances to be constrained at values $K^{(s)}$, s=1,...,S; and $V_1^{(s)}, V_2^{(s)}, V_3^{(s)}$ are variance model parameters for s=1,..., S.

Waters and Chester noted that there is no algebraic solution to this problem. They stated that a common approach to the problem is to examine the S univariate optima, which minimize C subject to a single variance equality constraint. If some of these univariate optima also satisfy the other variance inequalities, then the cheapest such solution is the solution to the general problem. If none of the univariate optima satisfy this criteria, then the authors suggest checking the $\frac{S(S-1)}{2}$ bivariate solutions, which minimize cost subject to two variance equality constraints. The cheapest such feasible solution is also the solution to the general problem. Waters and Chester suggest plotting lines of fixed variance on a two-way plot of m vs q. This graph is useful to suggest which of the $\frac{S(S-1)}{2}$ solutions is likely to be optimal, and also to highlight the most influential variance constraints. I discuss in Section 4 one alternative to this approach for the two variable case.

Deville(1993) provided several optimal clustering methods for more complex two-stage sample designs. He discussed the case of an unequal probability first stage sample with a simple random second stage sample, with a single constraint on the variance of a Horvitz-Thompson estimator. A sample-dependent cost model was formulated, and its design-expectation was derived. A super-population model was formulated, and the variance derived with respect both to the randomization and super-population measures. This resulted in a more tractable variance expression than the usual design-based variance.

Deville allowed the cluster size q_k to be different for all PSUs, unlike previous optimal clustering literature. However, it was found that all cluster sizes were equal at the optimum, and given by the standard formula (4) except where the optimal cluster size exceeds the population size within some PSUs, which would be rare in practice.

Deville also considered other situations, such as stratified second stage samples, and where auxiliary variables are available to assist both design and estimation. These situations are not discussed here, as this paper is focused on area-based sample design. The auxiliary information required by Deville's other methods would rarely be available for area-based sample design, although the methods would clearly be applicable in many cases, including the Census quality control study discussed by Deville.

3 Methods for Setting a Cluster Size for Several Area Types

This Section discusses optimal clustering methods where different cluster sizes can be chosen for each area type. It is assumed that the aim is to minimize total cost subject to a single overall variance constraint. Hansen, Hurwitz and Madow (1953) (vol.2) provide a method for this situation. It is assumed that cost and variance models have already been estimated for each area type, and are of the form:

(6) $$C = \sum_{i=1}^{A}(C_{1i} + C_{2i}m_i + C_{3i}m_iq_i)$$
(7) $$V = \sum_{i=1}^{A}\left(V_{1i} + V_{2i}m_i^{-1} + V_{3i}m_i^{-1}q_i^{-1}\right)$$

where: C_{1i} , C_{2i} , C_{3i} , V_{1i} , V_{2i} , V_{3i} are parameters for area type i, i=1, ..., A;
m_i = number of sample PSUs in area type i, i=1, ...,A;
q_i = cluster size in area type i, i=1,..., A;
$n_i = m_i q_i$ = number of final units in sample in area type i, i=1, ..., A.

The optimization problem can be written as:

(8) minimize C with respect to m_i , q_i (i=1, ..., A)
subject to V = K, for $q_i, m_i \in \mathbf{R}$

where K is a given constraint on the variance;
C and V are defined as in (6) and (7).

The solution to (8) is given by Hansen, Hurwitz and Madow (1953) (vol.2) is:

$$\textbf{(9)} \quad q_{i(opt)} = V_{2i}^{-\frac{1}{2}} V_{3i}^{\frac{1}{2}} C_{2i}^{\frac{1}{2}} C_{3i}^{-\frac{1}{2}}$$

$$m_{i(opt)} = \left(K - \sum_{j=1}^{A} V_{1j}\right)^{-1} \left(\sum_{j=1}^{A} V_{2j}^{\frac{1}{2}} C_{2j}^{\frac{1}{2}} + \sum_{j=1}^{A} V_{3j}^{\frac{1}{2}} C_{3j}^{\frac{1}{2}}\right) V_{2i}^{\frac{1}{2}} C_{2i}^{-\frac{1}{2}}$$

$$n_{i(opt)} = \left(K - \sum_{j=1}^{A} V_{1j}\right)^{-1} \left(\sum_{j=1}^{A} V_{2j}^{\frac{1}{2}} C_{2j}^{\frac{1}{2}} + \sum_{j=1}^{A} V_{3j}^{\frac{1}{2}} C_{3j}^{\frac{1}{2}}\right) V_{3i}^{\frac{1}{2}} C_{3i}^{-\frac{1}{2}}$$

As would be expected, (9) is equivalent to the single area type result (4), if A=1. Even for A≠ 1, $q_{i(opt)}$ in (9) is equivalent to result (4), however the $m_{i(opt)}$ and $n_{i(opt)}$ are different to result (4). Result (9) can be considered to be an allocation method taking into account cost, variance and the use of optimal cluster sizes. This method would not generally provide equally precise estimators for different area types, rather resources would be assigned optimally to area types to meet overall objectives. A common consequence would be higher relative variances for rural estimates than urban estimates. If some or all area type variances were of individual interest, then this could expressed as multiple variance constraints, discussed in Section 4.

In practice, both m_i and q_i must be positive integers. The distinction between real and integer optima is unimportant for m_i which would generally be large, however q_i is a much smaller value. It is suggested that an integer solution be obtained by evaluating the cost function at the 2^A neighboring integer solutions (that is ceilings and truncations of the A real optimal q_i). Each $\{m_i\}$ should also be recalculated at each of these 2^A possible solutions. It is simple to show that the optimal m_i for fixed q_i is given by:

$$\textbf{(10)} \quad m_i = \sqrt{\frac{\left(V_{2i}+V_{3i}q_i^{-1}\right)}{(C_{2i}+C_{3i}q_i)}} \frac{\sum_{j=1}^{A} \sqrt{(C_{2j}+C_{3j}q_j)(V_{2j}+V_{3j}q_j^{-1})}}{\left(K-\sum_{j=1}^{A} V_{1j}\right)}$$

(Proof available from author on request.)

This method would provide the exact integer optimum if the constrained optimization problem was everywhere convex in $\{m_i, q_i: i=1,...,A\}$. The problem is not necessarily convex everywhere but is convex at the optimum, so that the proposed method could be expected to equal to or be very close to the correct integer optimum.

In some situations, the sample sizes for each area type are constrained up to a single factor, that is:

$$\textbf{(11)} \quad n_i = \zeta n_{0i} \text{ for some } \zeta \in \mathbf{R},$$

where n_{0i} are fixed in advance. For example, the MLFS sample design included this constraint, as discussed in Section 5.4.

The solution to problem (8) with additional constraint (11) is given by:

$$\textbf{(12)} \quad q_{i(opt)} = \sqrt{\left(\frac{C_{2i}n_{0i}}{V_{2i}n_{0i}^{-1}}\right) \Big/ \left(\frac{\sum_{j=1}^{A} C_{3j}n_{0j}}{\sum_{j=1}^{A} V_{3j}n_{0j}^{-1}}\right)}$$

$$\zeta = \left(K - \sum_{i=1}^{A}\right)^{-1} \sum_{i=1}^{A} n_{0i}(V_{2i}q_{i(opt)} + V_{3i})$$

with the m_i and n_i determined by:

$n_i = \zeta n_{0i}$
$m_i = n_i q_i$.

(Proof available from author on request.)

4. Multivariate Optimal Clustering

In practice, there are usually several estimators whose variance is of interest. The general form of the multivariate optimization problem for multiple area types is:

(13) minimize C with respect to m_i , q_i (i=1,...,A) for $q_i, m_i \in \mathbf{R}$

subject to $V^{(s)} \le K^{(s)}$ for r = 1,...,S.

where:

$$V^{(s)} = \sum_{i=1}^{A} \left(V_{1i}^{(s)} + V_{1i}^{(s)} m_i^{-1} + V_{1i}^{(s)} m_i^{-1} q_i^{-1}\right);$$

$V^{(s)}$ are the variances to be constrained at values $K^{(s)}$, s=1,...,S;
$V_{1i}^{(s)}, V_{2i}^{(s)}, V_{3i}^{(s)}$ are variance model parameters, for i=1, ..., A; s=1, ..., S.

The solution to this problem is discussed in Waters and Chester (1987), for a single area type. A similar method could be adopted here, however their approach requires accurate specification of variance constraints, otherwise some constraints may be highly influential. A simpler solution which would cover many practical situations is to optimize based on a linear combination of the variances, with the linear coefficients set by the user. For example, the coefficients could be chosen by trying several possible values and perusing the resulting solutions. This was the approach taken

for the MLFS design. There is some discussion in Kish(1988) of these two approaches to multiple objective sample design.

The optimization problem can then be expressed as:

(14) minimize C with respect to m_i , q_i (i=1,...,A) for $q_i, m_i \in \mathbf{R}$

subject to $\sum_{s=1}^{S} w^{(s)} V^{(s)} \leq K$

where: $V^{(s)} = \sum_{i=1}^{A} \left(V_{1i}^{(s)} + V_{2i}^{(s)} m_i^{-1} + V_{3i}^{(s)} m_i^{-1} q_i^{-1} \right)$;

K is a constraint, possibly based on historical variances;

$\{w^{(s)} : s=1,...,S\}$ are set by the user such that $\sum_{s=1}^{S} w^{(s)} = 1$, to reflect the priority on each of the S variances.

This problem can be expressed in the same form as the univariate problems in Section 2, so the solutions are identical except that, in (9),

V_{1i} , V_{2i} and V_{3i}

are replaced by

$$\sum_{s=1}^{S} V_{1i}^{(s)} \text{ , } \sum_{s=1}^{S} V_{2i}^{(s)} \text{ and } \sum_{s=1}^{S} V_{3i}^{(s)}$$

respectively, for i=1,...,A.

5. Application to MLFS

5.1 Introduction

The MLFS is a monthly survey of about 30 000 dwellings, collecting information on labour force participation and related information. Two key estimates are the total number of employed persons and the total number of unemployed persons in Australia. The first stage is a probability proportional to size sample of primary sampling units (PSUs). The PSUs, usually referred to as collectors districts, contain about 300 dwellings on average and are a by-product of the five-yearly Census of Population and Housing. The second stage of sampling consists of a sample of dwellings from within each selected PSU. The cluster size refers to the number of dwellings to be selected from each selected PSU.

This is a slight simplification of the sample design, as there is in reality a third stage of sampling. PSUs are divided into blocks, and the final sample of dwellings is a systematic sample from a randomly selected block in each selected PSU. However, a two-stage design of dwellings within PSUs is considered to be an adequate approximation for the purpose of calculating optimal cluster sizes.

The MLFS sample is redesigned every five years and the work reported in this paper was part of the 1996 Redesign. A major feature of this Redesign was the introduction of telephone interviewing. Up to August 1996, the MLFS was enumerated by personal interviewing (PI) only. From August 1996 onwards, "warm telephone interviewing" (TI) was introduced, where new respondents were personally interviewed while the great majority of continuing respondents were interviewed by telephone. The introduction of computer assisted personal interviewing (CAPI) was also evaluated in 1996 but was not adopted. However, both CAPI and TI were used in some of the operational tests on which the cost models were based, so that the cost models incorporate some CAPI costs which were not relevant to the survey as adopted.

In the 1996 and past MLFS Redesigns, optimal clustering has been applied to multiple area types. There were 6 main area types, including rural and 5 types of urban areas.

5.2 Variance Modelling

Variance models were calculated by generating all possible samples based on Census data for various cluster sizes and sample sizes. True variances could then be calculated, under the assumption that the Census data and simulated design were appropriate. This was numerically feasible as CDs and clusters within CDs are selected systematically for the MLFS. Some approximations were made to the sample design in this process, however these are not discussed in this paper, which focuses on the optimal clustering method. In addition, variance models were based on the Horvitz-Thompson estimator, whereas a post-stratified estimator by age, sex and a geographic variable is used in reality. The impact of this approximation has not been measured, however it is expected to have relatively little impact on the ratio of the first and second stage variance parameters, implying little impact on optimal clustering.

A model of the form (7) in Section 3 was then fitted to the variances obtained from Census resampling, for several variables including employed persons and unemployed persons. The final variance model did not greatly change between the 1991 and 1996 MLFS Redesigns.

5.3 Cost Modelling

Cost models were based on a large dataset of workload costs for personal interviewing, and a dataset containing telephone interviewing costs for a subset of 80 of these workloads. These workload costs were from a six month trial of TI and CAPI, so that the TI costs also include costs associated with CAPI, even though CAPI was not adopted for the MLFS.

The cost of enumerating each workload was subdivided by the type of interviewer activity. Interviewer activities were assigned to either the C_{1i} , C_{2i} or C_{3i} parameters, for example travel between CDs was judged to depend on the number of CDs and hence assigned to the C_{2i} parameters. Cost models were fitted separately for each of these three components, and then combined into two models of the form (6) in Section 3: a TI cost model and a PI cost model. The reason for this approach was that a fitting a single regression model of the

form (6) was not possible, due to collinearity between the number of CDs and number of households.

Table 4: Comparison of $\frac{C_{2i}}{C_{3i}}$ in Two Cost Models

Area Type	1996 Personal Interviewing	1996 Telephone Interviewing
1	2.69	1.04
2	2.88	1.11
3	2.95	1.13
4	2.80	1.08
5	5.44	1.78
6	4.84	1.78

Table 4 contains a comparison of two cost models: the 1996 model for telephone interviewing; and the 1996 model for personal interviewing. From the last two columns of Table 4, it can be seen that telephone interviewing significantly reduced the relative contribution of CD-level costs, which would result in a less clustered optimal design.

5.4 Optimal Clustering Calculations

Variances for both employed persons and unemployed persons were explicitly included in the optimal clustering calculations. To include both variances, the method specified in (14) in Section 4 was used. To apply this method, we needed to choose a single weight parameter (denoted w_{emp}), to reflect the relative priorities of the two estimators. The quantity to be constrained would then be:

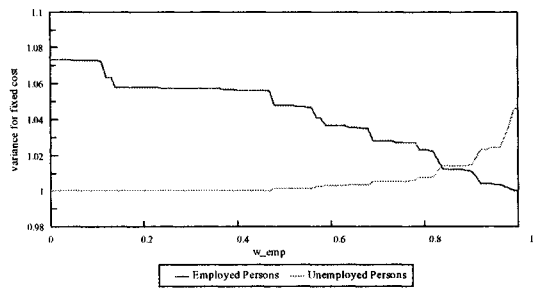
$$w_{emp}\,\text{var(employed persons)} + (1-w_{emp})\,\text{var(unemployed persons)}$$

One possible approach for choosing w_{emp} is to plot the solutions for different values of w_{emp} on a graph of employed persons variance against unemployed persons variance. However, for problem (8) the achieved cost is also different for different values of w_{emp}, so that it is difficult to weigh the different options. In order to remove the issue of cost, the converse problem was solved and plotted; that is the linear combination of variances was minimized for fixed cost. A small number of possible values of w_{emp} were chosen based on the plot, and the survey manager was provided with options based on these values.

Two values for w_{emp} were chosen based on Plot 1, which plots the employed and unemployed persons variances against w_{emp}, for a fixed cost constraint. Both variances have been scaled by dividing by the lowest possible variance for each variable, that is the variance obtained by setting w_{emp} to 0 for unemployed persons and setting w_{emp} to 1 for employed persons.

It can be seen from Plot 1 that setting w_{emp}=0.9 implies roughly equal priority on employed persons and unemployed persons, in the sense that both variances are above their minimum values by a factor of about 1.01. Unemployed persons has been given higher priority than employed persons in past redesigns. This could be reflected by setting w_{emp}=0.7, to retain some priority for employed persons, but to reduce the unemployed persons variance to close to its minimum possible value. Optimal clustering was calculated for both of these options, to evaluate which option better reflected current priorities.

Plot 1: Variance Tradeoff for Employment and Unemployment



There was an additional constraint on the design, which needed to be incorporated into optimal clustering calculations. Historically, MLFS sampling has been equal probability sampling (EPS) within each state. In addition, state sample sizes were initially set based on the priorities of state and national estimates; this was done prior to optimal clustering considerations. While the optimal clustering project could result in a change to overall sample size, this had to be done by increasing each state's sample size by a constant factor, to preserve the initial state sample relativities.

Clearly this places a constraint on the area type sample sizes. Let n_{0i} (i=1, ..., A) be the initial area type sample sizes, as implied by the initial state sample sizes and EPS within state. Let ζ be the factor to be applied to the initial state sample sizes. Then the final area type sample sizes must satisfy:

$$n_i = \zeta n_{0i} .$$

The optimal clustering problem was thus defined by (8) with constraint (11), as described in Section 3. Expression (12) from Section 3 gave a real-valued solution. The integer optimal cluster sizes were calculated by evaluating the cost function for fixed variance for all of the vectors of integers neighboring the real-valued optimal cluster sizes, as discussed in Section 3.

Table 5 summarizes two options for optimal clustering, for w_{emp}=0.7 and w_{emp}=0.9 respectively. The table gives the variances and the costs based on the TI cost model.

Option 2 was adopted, reflecting a small change in user priorities towards employment statistics.

Table 5: Two Optimal Clustering Options

	Option 1: w_{emp}=0.7	Option 2: w_{emp}=0.9
Change in Cost	-0.52%	-0.69%
Change in Sample Size	+0.15%	-2.0%
Change to Number of CDs	-3.25%	+8.04%
Standard Errors **Unemployed Persons** **Employed Persons**	 10 793 24 652	 10 849 24 499
Optimal Cluster Sizes		
1 Met Inner City	4	4
2 Met Inner Settled	7	6
3 Met Settled	8	7
4 Met Outer Growth	6	5
5 Ex-Met Urban	9	8
6 Ex-Met Rural	10	10

5.5 Gain from Splitting Cost and Variance Models by Area Type

The MLFS sample design is made more complex by allowing separate cost and variance models and cluster sizes in different area types. To measure whether this additional complexity gives sufficient efficiency gains, optimal clustering was calculated based on a single pooled cost and variance model. The cost of this design was then compared to the cost of a design based on models split by area type. Both the pooled and area type models were based on 1996 variance and TI cost models, and the cost of each was estimated using the 1996 TI cost model split by area type.

To simplify comparison, the state sample fraction constraint was not imposed, and non-integer cluster sizes were allowed, for both the pooled and optimal designs. The same variance constraint using w_{emp}=0.9 was imposed on both models. It was found that the cost for the optimal design split by area type was about 2.5% lower than the single area type design.

5.6 Effect of Telephone Interviewing on Optimal Clustering

Introducing telephone interviewing should result in greatly reduced travel costs, with relatively little impact on interview costs. As a result, the optimal TI design was expected to be less clustered than the optimal personal interviewing design. To explore this issue, optimal TI and PI designs were compared. Both optimal designs were based on the variance model with w_{emp}=0.9, and cost models calculated from current data for TI and PI. Mode of collection was assumed not to significantly affect variances.

The optimal PI design was much more clustered, with cluster sizes ranging from 6 to 22, compared to the optimal TI cluster sizes ranging from 4 to 10. Overall, the optimal TI design had 44% more PSUs and 9.2% lower sample size than the optimal PI design.

6 Summary

The existing literature includes methods for minimizing cost subject to one or more variance constraints. Most of the cost models assumed are linear, of the form (3) in Section 2, although some extensions to nonlinear models are given in Hansen, Hurwitz and Madow (1953). Section 3 summarized methods available for splitting models by area type, but still with overall cost and variance objectives. This results in the same optimal cluster size formula as in the existing literature, but a different allocation of sample to area types which takes account of optimal clustering and cost and variance. Optimal cluster sizes were derived for an additional constraint, where area type sample sizes are fixed up to a single factor. Section 4 briefly discussed some methods for dealing with multiple variance constraints.

Section 5 summarized the application of this theory to a Redesign of the Australian Monthly Labour Force Survey (MLFS), which has just introduced partial telephone interviewing (TI). For this survey, splitting cost and variance models by area type resulted in about 2.5% cost saving. It was found that the optimal TI sample was much less clustered than an optimal sample for personal interviewing, with about 10% less dwellings and 44% more PSUs. From this case, it can be concluded that splitting models by area type for optimal clustering is a valuable sample design tool, and that optimal clustering changes significantly if the mode of collection is changed.

7 References

(1) Brooks, S. (1955), "The estimation of an optimum subsampling number", Journal of the American Statistical Association 50, pp.398-415.

(2) Cochran (1977), Sampling Techniques, John Wiley and Sons New York.

(3) Deville (1993), "Optimum two-Stage sample Design for ratio estimators: application to quality control - 1990 French Census", Survey Methodology Vol.19, No. 2, pp. 173-182.

(4) Hansen, Hurwitz and Madow (1953), Sample Survey Methods and Theory, John Wiley and Sons New York.

(5) Kish (1988), "Multipurpose sample designs", Survey Methodology Vol.14, pp.19-32.

(6) Snedecor and Cochran (1980), Statistical Methods, Iowa State University Press.

(7) Waters and Chester (1987), "Optimum allocation in multivariate, two-stage sampling designs", The American Statistician Vol.41, No.1, pp. 46-50.